(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Patentschrift**

(10) **DE 197 36 669 C 1**

(51) Int. Cl.[6]:
G 10 L 5/00
H 03 M 7/30

(21) Aktenzeichen: 197 36 669.4-53
(22) Anmeldetag: 22. 8. 97
(43) Offenlegungstag: –
(45) Veröffentlichungstag der Patenterteilung: 22. 10. 98

Innerhalb von 3 Monaten nach Veröffentlichung der Erteilung kann Einspruch erhoben werden

(73) Patentinhaber:
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., 80636 München, DE

(74) Vertreter:
Schoppe, F., Dipl.-Ing.Univ., Pat.-Anw., 81479 München

(72) Erfinder:
Hilpert, Johannes, Dipl.-Ing., 90763 Fürth, DE; Herre, Jürgen, Dipl.-Ing., 91054 Buckenhof, DE; Grill, Bernhard, Dipl.-Ing., 91207 Lauf, DE; Buchta, Rainer, Dipl.-Ing., 91074 Herzogenaurach, DE; Brandenburg, Karl-Heinz, Dr.-Ing., 91054 Erlangen, DE; Gerhäuser, Heinz, Dr.-Ing., 91344 Waischenfeld, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

FR 27 41 743 A1

Zollner, M., Zwicker, E., Elektroakustik, Springer-Verlag, Berlin, Heidelberg, New York, 3. Auflage, 1993;

(54) Verfahren und Vorrichtung zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal sowie Vorrichtung und Verfahren zum Codieren eines Audiosignals

(57) Ein Verfahren zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal wird vollständig im Zeitbereich durchgeführt und umfaßt den Schritt des Segmentierens des zeitdiskreten Audiosignals, um aufeinanderfolgende Segmente gleicher Länge mit ungefilterten zeitdiskreten Audiosignalen zu erzeugen. Das zeitdiskrete Audiosignal in einem aktuellen Segment wird anschließend gefiltert. Nun kann entweder die Energie des gefilterten zeitdiskreten Audiosignals in dem aktuellen Segment mit der Energie des gefilterten zeitdiskreten Audiosignals in einem vorhergehenden Segment verglichen werden, oder es kann ein aktuelles Verhältnis zwischen der Energie des gefilterten zeitdiskreten Audiosignals in dem aktuellen Segment und der Energie des ungefilterten zeitdiskreten Audiosignals in dem aktuellen Segment gebildet werden und dieses aktuelle Verhältnis mit einem vorhergehenden entsprechenden Verhältnis verglichen werden. Auf der Grundlage des einen Vergleichs und/oder des anderen Vergleichs wird erfaßt, ob in dem zeitdiskreten Audiosignal ein Anschlag vorhanden ist.

DE 197 36 669 C 1

DE 197 36 669 C 1

Beschreibung

Die vorliegende Erfindung bezieht sich auf die Codierung von Audiosignalen und insbesondere auf das Codieren von Audiosignalen, die Anschläge aufweisen, d. h. die transient sind.

Bei der gehörangepaßten Codierung zur Datenreduktion von Audiosignalen erfolgt die Codierung der Audiosignale zumeist im Frequenzbereich. Dies bedeutet, daß Ausgangswerte einer Zeit-Frequenz-Transformation quantisiert werden und anschließend in einen Bitstrom geschrieben werden, welcher gespeichert oder übertragen werden kann. Ein psychoakustisches Modell, das in dem Codierer implementiert ist, berechnet eine momentane Mithör- oder Maskierungsschwelle und steuert die Quantisierung der Ausgangswerte der Zeit-Frequenz-Transformation so, daß der Codierungsfehler, d. h. der Quantisierungsfehler, spektral geformt wird und unter dieser Schwelle liegt, damit derselbe unhörbar ist. Durch diese Maßnahme ist der Codierungsfehler jedoch zeitlich über der Zahl von Abtastwerten konstant, welche der Länge des Transformationsfensters entspricht. Die Mithör- oder Maskierungsschwelle ist in M. Zollner, E. Zwicker, Elektroakustik, Springer-Verlag, Berlin, Heidelberg, New York, 3. Auflage, 1993 dargestellt.

Um die Berechnung der Mithörschwelle im Frequenzbereich möglichst exakt durchführen zu können, ist eine hohe Frequenzauflösung der Zeit-Frequenz-Transformation erforderlich. Bei praktischen Anwendungsfällen können typische Transformationslängen im Bereich von 20 bis 40 ms auftreten. Werden nun transiente Audiosignalen, d. h. Audiosignale mit Anschlägen, verarbeitet, so verteilt sich je nach zeitlicher Position des Anschlags im Transformationsfenster das Quantisierungsgeräusch zeitlich auch "vor" das Maximum der Signalhüllkurve. Aufgrund der menschlichen Wahrnehmung können diese sogenannten "Vorechos" hörbar werden, wenn sie mehr als 2 ms vor dem eigentlichen Anschlag des zu codierenden Audiosignals einsetzen. Dies ist der Grund, daß bei vielen Transformationscodierern die Transformationslänge der Zeit- Frequenz-Transformation auf kürzere Fenster, d. h. kürzere Blocklängen, mit einer zeitlichen Länge von typischerweise 5 bis 8 ms mit einer damit höheren zeitlichen Auflösung umschaltbar ist. Dies ermöglicht eine zeitlich feinere Formung des Quantisierungsgeräusches und damit eine Unterdrückung dieser Vorechos, wodurch dieselben nicht mehr oder nur sehr wenig hörbar sind, wenn das codierten Signal wieder in einem Decodierer decodiert wird.

Es werden also Vorrichtungen zum Erfassen eines Anschlags in einem Audiosignal verwendet, um die Transformationslänge der Zeit-Frequenz-Transformation gehörrichtig an die Eigenschaften und insbesondere an die transienten Eigenschaften des Audiosignals anzupassen.

Fig. 3 zeigt einen bekannten Transformationscodierer **100**, der allgemein nach dem Standard MPEG 1--2 Layer3 (ISO/IEC IS 11172-3, Coding of Moving Pictures and Associated Audio, Part 3: Audio) aufgebaut ist. Ein Zeitsignal gelangt über einen Eingang **102** in einen Block Zeit/Frequenz-Transformation **104**. Das Zeitsignal am Eingang **102**, das typischerweise als zeitdiskretes Audiosignal vorliegt, das mittels einer Abtasteinrichtung (nicht gezeigt) aus einem zeitkontinuierlichen Zeitsignal erhalten wurde, wird durch den Block Zeit-Frequenz -Transformation **104** in aufeinanderfolgende Blöcke von Spektralwerten transformiert, welche in einen Block Quantisierung/Codierung **106** eingegeben werden, wobei das Ausgangssignal des Blocks Quantisierung/Codierung quantisierte und Redundanz-codierte digitale Signale sind, welche in einem Block Bitstromformatierung **108** zusammen mit nötigen Seiteninformationen zu einem Bitstrom gebildet werden, der am Ausgang der Bitstromformatierungseinrichtung **108** anliegt und gespeichert oder übertragen werden kann.

In dem Block Zeit/Frequenz-Transformation **104** findet eine Fensterung des zeitdiskreten Audiosignals am Eingang **102** statt, um aufeinanderfolgende Blöcke mit zeitdiskreten Audiosignalen, welche nun gefenstert sind, zu erzeugen. Die Blöcke der gefensterten zeitdiskreten Audiosignale werden anschließend, wie es bereits erwähnt wurde, in den Frequenzbereich transformiert. Wie es aus der Nachrichtentechnik bekannt ist, ist die Frequenzauflösung der Zeit-Frequenz-Transformation durch die Länge eines Blocks vorgegeben. Um für zeitdiskrete Audiosignale mit Anschlägen, d. h. mit transienten Anteilen, eine ausreichende zeitliche Auflösung zu erreichen, ist es notwendig, daß zur Codierung derselben zur Vermeidung der Vorechos die Fensterlänge und damit die zeitliche Länge eines Blocks zeitdiskreter Abtastwerte verkürzt wird.

Der in **Fig.** 3 gezeigte bekannte Codierer führt folgendes Verfahren zum Erfassen von Anschlägen in einem Audiosignal durch. Aus dem Block Zeit/Frequenz-Transformation **104** werden die Spektralkomponenten in einen Block psychoakustisches Modell **110** eingespeist, wobei der Block **110** zum einen, wie es bereits eingangs erwähnt wurde, die Maskierungs- oder Mithörschwelle für den Block Quantisierung/Codierung **106** ermittelt, sowie zum anderen aus dem vorliegenden Signalenergieverlauf des zeitdiskreten Audiosignals im Frequenzbereich und dem errechneten Energieverlauf der Mithörschwelle einen Schätzwert für den Bitbedarf zur Codierung des Spektrums ermittelt. Der geschätzte Bitbedarf, der in der Fachwelt auch "Perceptual Entropy" oder kurz "pe" genannt wird, berechnet sich aus folgendem Zusammenhang:

$$pe = \sum_{k=1}^{N} \frac{1}{2} \log_2 \left(\frac{e(k)}{n(k)} + 1 \right) \qquad (1)$$

In Gleichung (1) bedeuten N die Anzahl der Spektrallinien eines Blocks, e(k) die Signalenergie der Spektralkomponenten oder Spektrallinien k und n(k) die erlaubte Störenergie der Linie k. Ein Anstieg dieser Perceptual Entropy von einem Transformationsfenster zum nächsten, welcher einen gewissen Schwellenwert, der als "switch_pe" bezeichnet wird, übersteigt, dient hier als Indikator für einen Anschlag. Wird der Schwellenwert switch_pe überschritten, so wird in dem Block **104** von einem langen Fenster zu einem kurzen Fenster umgeschaltet, um zeitlich kürzere Blöcke zeitdiskreter Audiosignale zu erzeugen, um die Zeitauflösung des Transformationscodierers **100** zu erhöhen. Die Berechnungsvorschrift, die in Gleichung (1) dargelegt ist, sowie die Festlegung des Schwellenwerts switch_pe werden in einem Block Bitbedarfsschätzung **112** festgelegt. Das Ergebnis der Bitbedarfsschätzung **112** wird der Zeit/Frequenz-Transformation **104** sowie dem psychoakustischen Modell **110** mitgeteilt, wie es in **Fig.** 3 angedeutet ist.

Ein Nachteil dieses bekannten Verfahrens besteht darin, daß die Informationen über einen möglichen Anschlag oder

"Attack" erst nach der Berechnung des psychoakustischen Modells zur Verfügung stehen. Dies wirkt sich insbesondere nachteilig auf die zeitliche Ablaufstruktur des Codierers aus, da eine Rückkopplung der Fensterinformationen zum psychoakustischen Modell erfolgen muß. Weiterhin wirken sich Änderungen an Parametern zur Berechnung der Mithörschwelle immer auch auf den Wert der Perceptual Entropy aus. Veränderungen dieser Parameter verändern daher immer auch die Fenstersequenz, d. h. die Folge von langen und kurzen Fenstern, der Transformation.

Fig. 4 zeigt einen weiteren bekannten Transformationscodierer **150**, der im prinzipiellen Aufbau dem Transformationscodierer **100** ähnelt. Insbesondere umfaßt derselbe ebenfalls den Eingang **102** für zeitdiskrete Audiosignale, welche in dem Block **104** gefenstert und in den Frequenzbereich transformiert werden. In dem Block **106** werden die spektralen Ausgangswerte des Blocks **104** unter Berücksichtigung des psychoakustischen Modells **110** quantisiert und anschließend codiert, um zusammen mit Seiteninformationen durch die Bitstromformatierungseinrichtung **108** in einen ausgangsseitigen Bitstrom geschrieben zu werden.

Der Transformationscodierer **150**, der in **Fig. 4** gezeigt ist, unterscheidet sich von dem Transformationscodierer **100**, der in **Fig. 3** gezeigt ist, in der Erfassung von Anschlägen in dem Audiosignal. Die Erfassung von Anschlägen in dem Audiosignal am Eingang **102**, die in **Fig. 4** dargestellt ist, ist in dem Standard MPEG 2 AAC (siehe ISO/IEC IS 13818-7, MPEG-2 Advanced Audio Coding (AAC)) beschrieben. Der Block FFT-Transformation und Detektion aus dem Spektrum **152** führt eine Erfassung von Anschlägen mittels eines spektralen Energieanstiegs durch. Insbesondere wird das zeitdiskrete Audiosignal am Eingang **102** zuerst mittels einer FFT-Transformation in den Frequenzbereich transformiert, wobei die Länge der FFT-Transformation der Transformationslänge der kurzen Fenster entspricht. Anschließend werden die FFT-Energien in den sogenannten "Critical Bands" berechnet. Die "Critical Bands" stellen eine Frequenzgruppierung dar, die der Auflösung des psychoakustischen Modells entspricht. Ein Schwellwertvergleich der einzelnen Bandenergien über eines oder mehrere zeitlich aufeinanderfolgende Fenster liefert nun ein Indiz für einen Anschlag.

Im Gegensatz zu dem in **Fig. 3** dargestellten bekannten Verfahren vermeidet das in **Fig. 4** dargestellte bekannte Verfahren den Nachteil der Rückkopplung der Fensterinformationen zum psychoakustichen Modell **110**. Das in **Fig. 4** dargestellte Verfahren könnte prinzipiell unabhängig vom psychoakustischen Modell vor dessen Berechnung angewendet werden. Das in **Fig. 4** dargestellte Verfahren benötigt jedoch eine an die Transformationslänge des Codierers angepaßte FFT-Transformation zur Berechnung der Energien in den einzelnen Frequenzgruppen.

Aus der FR 2741743 A1 ist bereits ein Verfahren zum Verbessern der Erkennbarkeit eines Wortes in einem Sprachverschlüsselungsgerät mit niedriger Datenrate bekannt, bei dem ein Mustersprachsignal in Rahmen von vorgegebener Zeitdauer zerlegt wird, wobei jedem Rahmen ein Prädiktionsfilter zugeordnet wird, dessen Koeffizienten unter Berücksichtigung der Anordnung stabiler oder instabiler Mustersignale ermittelt sind. Zum Zwecke der Bestimmung der stabilen oder unstabilen Eigenschaft des Mustersignals wird jeder Rahmen in eine vorbestimmte Anzahl von Teilrahmen zerlegt, die Leistung des Signals in jedem Teilrahmen ermittelt und eine Anzahl von Leistungsmessungen in jedem Teilrahmen für das verbliebene Signal am Ausgang des Prädiktionsfilters während eines Zeitfensters vorgenommen, wobei die erhaltenen Leistungsmessungen mit der Mustersignalenergie des entsprechenden Teilrahmens verglichen wird.

Die Aufgabe der vorliegenden Erfindung besteht darin, ein Verfahren und eine Vorrichtung zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal sowie ein Verfahren und eine Vorrichtung zum Codieren von Audiosignalen zu schaffen, welche auf effiziente und einfache Art und Weise eine zuverlässige Erfassung von Anschlägen und somit eine einfache Unterdrückung von Vorechos ermöglichen.

Diese Aufgabe wird durch ein Verfahren zum Erfassen eines Anschlags gemäß Anspruch 1, durch eine Vorrichtung zum Erfassen eines Anschlags gemäß Anspruch 11, durch eine Vorrichtung zum Codieren eines zeitdiskreten Audiosignals gemäß Anspruch 14 und durch ein Verfahren zum Codieren eines zeitdiskreten Audiosignals gemäß Anspruch 15 gelöst.

Der vorliegenden Erfindung liegt die Erkenntnis zugrunde, daß ein Anschlag in einem Audiosignal mit einem zeitlichen Anstieg der Signalenergie des Audiosignals einhergeht. Ferner löst ein Anschlag einen Anstieg der Energie von höherfrequenten Signalanteilen in dem Audiosignal aus, da ein Anschlag typischerweise durch schnelle zeitliche Änderungen des Audiosignals gekennzeichnet ist.

Ein Verfahren zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal umfaßt somit folgende Schritte:

(a) Segmentieren des zeitdiskreten Audiosignals, um aufeinanderfolgende Segmente gleicher Länge mit ungefilterten zeitdiskreten Audiosignalen zu erzeugen;
(b) Filtern des zeitdiskreten Audiosignals in einem aktuellen Segment;
(c) Vergleichen der Energie des gefilterten zeitdiskreten Audiosignals in dem aktuellen Segment mit der Energie des gefilterten zeitdiskreten Audiosignals in einem vorhergehenden Segment; und/oder
(d) Bestimmen eines aktuellen Verhältnisses zwischen der Energie des gefilterten zeitdiskreten Audiosignals in dem aktuellen Segment und der Energie des ungefilterten zeitdiskreten Audiosignals in dem aktuellen Segment und Vergleichen des aktuellen Verhältnisses mit einem entsprechenden vorhergehenden Verhältnis; und
(e) Erfassen eines Anschlags auf der Grundlage des im Schritt (c) und/oder im Schritt (d) geführten Vergleichs.

Bei einem bevorzugten Ausführungsbeispiel wird das Filtern mittels eines Hochpaßfilters ausgeführt, es sind jedoch auch andere Filterungen möglich, z. B. mittels eines Bandpaßfilters, eines Differenzierers erster oder höherer Ordnung oder ähnlichem, solange sich das gefilterte zeitdiskrete Audiosignal hinsichtlich seiner spektralen Eigenschaften von dem ungefilterten zeitdiskreten Audiosignal unterscheidet.

Der in dem Schritt (c) durchgeführte Vergleich dient zum Erfassen eines zeitlichen Anstiegs der Signalenergie, d. h. zur Anstiegsdetektion, während der in dem Schritt (d) durchgeführte Vergleich zur Erfassung des Anstiegs von Signalanteilen eines bestimmten Frequenzbereichs, d. h. zur Spektraldetektion, dient.

Der in dem Schritt (d) durchgeführte Vergleich dient dagegen dazu, frequenzabhängige Effekte der zeitlichen Maskierung zu berücksichtigen. An dieser Stelle sei angemerkt, daß die Zeitauflösung des menschlichen Ohrs frequenzabhängig ist. Die Zeitauflösung ist grob gesprochen bei sehr niedrigen Frequenzen relativ gering und steigt mit zunehmender Fre-

quenz an. Für den Fall eines Vorechos bedeutet dies, daß ein durch die Quantisierung eingeführtes Rauschen, das zu einem Vorecho in einem bestimmten zeitlichen Abstand vor einem Anschlag führt, bei niedrigen Frequenzen eher nicht erfaßt wird, da das Ohr hier eine zeitliche Auflösung hat, die gröber als der bestimmte zeitliche Abstand des Vorechos ist. Anders geartet ist der Fall, wenn ein Anschlag eher im höherfrequenten Bereich stattfindet. Hier ist die Zeitauflösung des menschlichen Ohres feiner, wodurch ein Vorecho in dem bestimmten zeitlichen Abstand hörbar sein kann, da die Zeitauflösung des Ohrs bereits feiner als der zeitliche Abstand des Vorechos vom Anschlag sein kann. Es bleibt also festzustellen, daß die Spektraldetektion im Gegensatz zur Anstiegsdetektion die frequenzabhängige zeitliche Auflösung des Ohrs nachbildet, wodurch eine genauere Anschlagserfassung als mit der Anschlagsdetektion allein möglich ist. In manchen Fällen kann selbstverständlich auch die Anschlagsdetektion alleine bereits zufriedenstellende Ergebnisse liefern.

An dieser Stelle sei angemerkt, daß ein Anschlag entweder auf der Grundlage des in dem Schritt (c) durchgeführten Vergleichs oder auf der Grundlage des in dem Schritt (d) durchgeführten Vergleichs oder auf der Grundlage beider Vergleiche durchgeführt werden kann.

Bevorzugte Ausführungsbeispiele der vorliegenden Erfindung werden nachfolgend bezugnehmend auf die beiliegenden Zeichnungen detaillierter erläutert. Es zeigen:

Fig. 1 einen Transformationscodierer, der die Anschlagserfassung im Zeitbereich umfaßt;

Fig. 2 eine detailliertere Darstellung der in **Fig.** 1 enthaltenen Anschlagserfassung im Zeitbereich;

Fig. 3 einen Transformationscodierer, der ein bekanntes Verfahren zur Anschlagserfassung umfaßt; und

Fig. 4 einen weiteren Transformationscodierer, der ein anderes bekanntes Verfahren zur Anschlagserfassung aufweist.

Fig. 1 zeigt einen Transformationscodierer **10** gemäß der vorliegenden Erfindung, welcher sich bis auf einen Block Anschlagserfassung **12** nicht von üblichen in der Technik bekannten Transformationscodierern unterscheidet. Insbesondere sind die Funktionen und Verknüpfungen der Blöcke Zeit/Frequenz-Transformation **104**, Quantisierung/Codierung **106**, Bitstromformatierung **108** und psychoakustisches Modell **110** in der Technik bekannt. Die Funktionsweisen der einzelnen Blöcke wurden bereits in Verbindung mit den **Fig.** 3 und 4 beschrieben und werden daher nicht noch einmal explizit erklärt.

Wie es in **Fig.** 1 gezeigt ist, erhält der Block Anschlagserfassung **12** als Eingangssignal das zeitdiskrete Audiosignal über den Eingang **102** des Transformationscodierers **10**. Der Block Anschlagserfassung **12** liefert als Ausgangssignal ein Signal, das anzeigt, ob ein langes oder kurzes Fenster für die Fensterung und anschließende Zeit/Frequenz-Transformation **104** festzulegen ist.

Fig. 2 zeigt eine detaillierte Ansicht des Blocks Anschlagserfassung **12** von **Fig.** 1. Das zeitdiskrete Audiosignal x(k), das an dem Ausgang **102** des Transformationscodierers **10** (**Fig.** 1) anliegt, wird in eine Segmentierungseinrichtung **14** eingespeist, welche am Ausgang aufeinanderfolgende Segmente der Länge S ausgibt. Ein Segment umfaßt daher die Anzahl S von zeitdiskreten Abtastwerten des Audiosignals und wird als xs(T) bezeichnet, wobei "T" darstellt, daß es sich beim Signal xs(T) um das aktuelle Segment handelt, während "T-1" anzeigt, daß es sich um ein dem aktuellen Segment zeitlich unmittelbar vorausgehendes Segment handelt. "T-2" bedeutet analog, daß das Segment mit "T-2" das zweitletzte Segment vor dem aktuellen Segment ist.

Das Signal xs(T) wird ferner in ein Hochpaßfilter **16** einerseits sowie in eine Spektraldetektionseinrichtung **18** andererseits eingespeist. Das Ausgangssignal ys(T) des Hochpaßfilters **16** wird wiederum zum einen in eine Anstiegsdetektionseinrichtung **20** einerseits und in die Spektraldetektionseinrichtung **18** andererseits eingespeist. Das Ausgangssignal der Anstiegsdetektionseinrichtung **20** wird ebenso wie das Ausgangssignal der Spektraldetektionseinrichtung **18** einer Anschlagserfassungseinrichtung **22** zugeführt, welche als ODER-Gatter ausgeführt sein kann, wie es durch das Symbol "v" symbolisch in **Fig.** 2 gezeigt ist. Das Ausgangssignal der Anschlagserfassungsvorrichtung **22** entspricht dem Ausgangssignal der Anschlagserfassungseinrichtung **12** von **Fig.** 1 und wird dem Block Zeit/Frequenz-Transformation **104** sowie dem Block psychoakustisches Modell **110** zur Verfügung gestellt.

Im Nachfolgenden wird auf die Funktion und den Aufbau der einzelnen in **Fig.** 2 gezeigten Elemente eingegangen.

Die Segmentierungseinrichtung **14** teilt das Eingangssignal x(k) in aufeinanderfolgende Segmente $x_s(T)$, $x_s(T-1)$, $x_s(T-2)$, ... gleicher Länge S ein. Das zeitdiskrete Audiosignal $x_s(T)$ in einem aktuellen Segment (T) umfaßt somit S zeitdiskrete Abtastwerte des zeitdiskreten Audiosignals x(k) am Eingang **102**, wobei die Segmentlänge S unabhängig von der Blocklänge der Zeit/Frequenz-Transformation gewählt werden kann. Insbesondere ist es im Gegensatz zum Stand der Technik nicht erforderlich, als Segmentlänge z. B. die kurze Blocklänge oder die lange Blocklänge zu wählen. Die Segmentlänge S kann im Bereich von 200 bis zu 2000 Abtastwerten liegen, wobei eine Segmentlänge S von etwa 500 Abtastwerten bevorzugt wird.

Das Hochpaßfilter **16** erfüllt im wesentlichen zwei Aufgaben. Die Anstiegsdetektion (Block **20**) soll einen Anstieg in der Hüllkurve der Signalenergie detektieren, nicht jedoch dem Amplitudenverlauf eines tieffrequenten Signales folgen. Liegt nun die Schwingungsdauer eines Signalanteils in der Größenordnung der Segmentlänge oder darüber, würde unter Umständen eine Fehldetektion eines Anschlags erfolgen. Der Frequenzgang des Hochpaßfilters **16** sollte somit vorzugsweise eine genügende Sperrdämpfung im unteren Frequenzbereich besitzen. Mit zunehmender Sequenzlänge S kann zudem die Grenzfrequenz des Filters weiter verringert werden. Andererseits werden die Energien des hochpaßgefilterten Zeitsignals $y_s(T)$ weiterhin als Vergleichsmaß für die Spektraldetektion (Block **18**) benötigt.

Bezüglich der Flankensteilheit und Welligkeit im Durchlaßbereich kann das Filter sehr mäßige Eigenschaften aufweisen, wobei jedoch ein lineares Phasenverhalten bevorzugt wird. Bei einer bevorzugten Segmentlänge von etwa 500 Abtastwerten wird bei einem bevorzugten Ausführungsbeispiel der vorliegenden Erfindung ein nicht rekursives, linearphasiges FIR-Filter der Länge 7 mit den Filterkoeffizienten –0,2136, –0,0257, –0,0265, –0,5713, –0,0265, –0,0257, –0,2136 verwendet werden. Die Länge des FIR-Filters des bevorzugten Ausführungsbeispiel ist jedoch nicht auf den genannten Wert eingeschränkt. Für manche Fälle dürften auch Filter mit geringerer Länge ausreichen, während in wieder anderen Fällen deutlich mehr Filterkoeffizienten erwünscht sein könnten.

Weiterhin wird bevorzugt, daß die Filterlänge klein gegenüber der Segmentlänge S ist. In diesem Fall kann nämlich die Filterlaufzeit vernachlässigt werden, wodurch eine weitere Komplexität des Transformationscodierers **10** vermieden werden kann.

Die Segmente werden durch ein nicht rekursives, digitales Filter, wie es bereits erwähnt wurde, mit einer sehr kurzen Filterlänge im Vergleich zur Segmentlänge von tieffrequenten Anteilen befreit. Für die Ausgangsfolge des Filters $y_s(T)$ ergibt sich folgende Gleichung:

$$y_s(T) = x_s(T) \times h(k) \quad (2)$$

h(k) stellt in Gl. (2) die Impulsantwort des Filters dar, während k der Filterlänge entspricht. Das Ausgangssignal $y_s(T)$ entsteht also aus der Faltung des Eingangssignals $x_s(T)$ mit der Impulsantwort h(k) des Hochpaßfilters **16**.

In der Anstiegsdetektionseinrichtung **20** wird zunächst aus den gefilterten Eingangsdaten $y_s(T)$ über ein Skalarprodukt die Energie $E_f(T)$ des gerade vorliegenden Segmentes, das auch als aktuelles Segment bezeichnet wird, berechnet. Ein Vergleich mit der Energie $E_f(T-1)$ des dem aktuellen Segment vorausgehenden Segments sowie mit der Energie $E_f(T-2)$ des zweitletzten vorausgehenden Segments liefert nun das Kriterium für den Energieanstieg in dem zeitdiskreten Audiosignal von einem Segment zum nächsten. Der Ausdruck für das erste Kriterium oder kritA lautet somit folgendermaßen:

$$kritA = [E_f(T) > k_1 \cdot E_f(T-1)] \wedge [E_f(T) > k_2 \cdot E_f(T-2)] \wedge [E_f(T) > E_{minF}] \quad (3)$$

Entsprechend der üblichen Notation bedeutet "v" eine logische ODER-Verknüpfung während " $\wedge$ " eine logische UND-Verknüpfung bezeichnet. Der letzte Term der Gleichung 3 bezeichnet einen Vergleich der aktuellen Energie des tiefpaßgefilterten, zeitdiskreten Audiosignals in dem aktuellen Segment mit einer Filter-Mindestenergie E_{minF}. Dieser Vergleich bewirkt, daß das Kriterium A nur berücksichtigt wird, wenn die aktuelle Segmentenergie eine Mindestenergie überschreitet. Der Wert der Konstanten E_{minF} kann vorher festgelegt werden und basiert in vereinfachter Form auf dem Einfluß der Ruhehörschwelle auf die Wahrnehmung. Die Mindestenergie für den konstanten Wert E_{minF} kann daher vorzugsweise im Bereich von –80 dBFs liegen.

Die in dem Block **18** ausgeführte Spektraldetektion basiert dagegen auf einem Vergleich von gefilterten und ungefilterten Segmentenergien des aktuellen Segments mit gefilterten und ungefilterten Segmentenergien des vorhergehenden Segments. In Gleichungsform ausgedrückt ergibt sich folgende Vorschrift für das zweite Kriterium kritB:

$$kritB = \left[\frac{E_f(T)}{E_u(T)} > k_3 \cdot \frac{E_f(T-1)}{E_u(T-1)}\right] \wedge [E_u(T) > E_{minU}] \quad (4)$$

In dieser Gleichung stellt $E_u(T)$ die Energie des aktuellen ungefilterten Segments dar, während $E_f(T)$ die Energie des hochpaßgefilterten aktuellen Segments, d. h. die Energie des hochpaßgefilterten zeitdiskreten Audiosignals im aktuellen Segment, darstellt. Der letzte Term der Gleichung (4) berücksichtigt wieder den Fall, daß keine Fensterumschaltung ausgelöst wird, wenn die Energie des ungefilterten zeitdiskreten Audiosignals im aktuellen Segment unter einer Minimalenergie E_{minU} für ungefilterte Signale liegt, welche wiederum auf der Ruhehörschwelle basiert und ebenso wie die Filter-Minimalenergie E_{minF} einen Wert von –80 dBFs annehmen kann.

In den Gleichungen (3) und (4) sind ferner verschiedene Konstanten k_1 bis k_3 genannt. Mittels dieser Konstanten wird festgelegt, wieviel größer die Energie des aktuellen Segments bzw. das aktuelle Verhältnis zwischen gefilterter Energie und ungefilterter Energie im Vergleich zu dem entsprechenden Wert des vorausgehenden Segmente sein muß, damit ein Anschlag erfaßt wird, durch den eine Fensterumschaltung von langen zu kurzen Fenstern bewirkt wird.

In der Praxis haben sich Werte für die Konstanten k_1 und k_3 von vier als günstig erwiesen, welche damit einem entsprechenden Pegelunterschied von 6 dB entsprechen. Lediglich vorzugsweise kann die Konstante k_2, also der Vergleichswert mit der vorletzten Segmentenergie, auch etwas kleiner als vier gewählt werden, um beispielsweise einen Wert von drei anzunehmen. Es wir jedoch darauf hingewiesen, daß die Werte für die Konstanten k_1 bis k_3 abweichend von den genannten Werten eingestellt werden können, wenn eine feinere bzw. gröbere Anschlagserfassung gewünscht wird. Für eine korrekte Funktionsweise der Anschlagserfassung der vorliegenden Erfindung ist es jedoch erforderlich, daß die Werte der Konstanten k_1 bis k_3 größer als eins eingestellt werden, wie es aus den Gleichungen (3) und (4) ersichtlich ist.

An dieser Stelle sei angemerkt, daß das Kriterium A (kritA) und das Kriterium B (kritB) lediglich auf dem jeweils ersten Term der Gleichung (3) und (4) basieren können. Die weiteren beiden Terme in der Gleichung (3) sowie der weitere Term in der Gleichung (4) dienen lediglich einer ausgefeilteren Anschlagserfassung, um sicherzustellen, daß möglichst wenig Anschläge erfaßt werden, um möglichst selten zu den kurzen Transformationsfenstern umschalten zu müssen.

Um den Einfluß von Schwebungen auf die Anstiegsdetektion zu minimieren, ist der Vergleich der gefilterten Energien nicht nur mit der zeitlich vorhergehenden Segmentenergie $E_f(T-1)$ sondern zusätzlich mit dem vorletzten Energiewert $E_f(T-2)$ bei der gewählten Segmentlänge wünschenswert. Hier wird der Effekt der zeitlichen Nachverdeckung bei kurz aufeinanderfolgenden Anschlägen berücksichtigt, wenn ein potentielles Vorecho vor einem zweiten Anschlag noch vom ersten Anschlag maskiert wird. Der zweite Term in Gleichung (3) stellt für die Funktion der vorliegenden Erfindung keinen wesentlich Term dar, sondern lediglich eine vorteilhafte Ausgestaltung. Dasselbe trifft für die jeweils letzten Terme der Gleichungen (3) und (4) zu, welche das Erfassen eines Anschlags von Mindestenergien abhängig machen, die der Ruheschwelle nachempfunden werden.

An dieser Stelle sei noch einmal betont, daß die Verwendung des Hochpaßfilters lediglich beispielhaft wenn auch bevorzugt ist. Anstelle des Hochpaßfilters könnte genausogut ein Differenzierer eingesetzt werden, der allgemein ausgedrückt dazu führt, daß im differenzierten Signal höherfrequente Signalanteile stärker zutage treten als im nicht-differenzierten Signal. Eine weitere Alternative für das Hochpaßfilter wäre ein Bandpaßfilter, das dazu führt, daß die Energie des bandpaßgefilterten Signals in einem bestimmten Spektralbereich konzentriert ist. Diese Aufzählung der Alternativen für das Hochpaßfilter des bevorzugten Ausführungsbeispiels ist jedoch nicht erschöpfend. Voraussetzung für das Verfahren der vorliegenden Erfindung ist, daß das Signal im Zeitbereich verarbeitet, d. h. gefiltert wird, und zwar derart, daß es sich

hinsichtlich seiner spektralen Eigenschaften von dem nicht verarbeiteten, d. h. ungefilterten Signal unterscheidet. Der Ausdruck "Filtern" ist daher nicht derart begrenzend aufzufassen, daß der lediglich eine übliche Filterung z. B. mittels eines Hochpasses umfaßt, sondern daß er auch andere Verarbeitungen, wie z. B. Differenzierungen, umfaßt, die dazu führen, daß sich das verarbeitete Signal hinsichtlich seiner spektralen Eigenschaften von dem nicht verarbeiteten unterscheidet.

Weiterhin sei darauf hingewiesen, daß die Einrichtung **22** zum Erfassen eines Anschlags nicht unbedingt als ODER-Gatter ausgeführt sein muß. Dieselbe kann z. B. als UND-Gatter ausgeführt sein. In diesem Fall wird nur dann ein Anschlag erfaßt, wenn beide Kriterien erfüllt sind. In diesem Fall würden vorzugsweise die Konstanten k_1, k_2 und/oder k_3 und/oder die Mindestenergien verkleinert werden, was dazu führt, daß jedes Kriterien für sich einfacher erfüllt wird. Um jedoch keine unnötigen oder zu häufigen Umschaltungen auf kürzere Fenster zu bewirken, wird dann ein Anschlag nur erfaßt, wenn beide Kriterien in einem Segment gleichzeitig erfaßt werden.

Die vorliegende Erfindung schafft somit eine Detektion von Anschlägen in Audiosignalen aus der Zeitsignalfolge, welche ausschließlich im Zeitbereich stattfindet. Die Anschlagserfassung bietet daher gegenüber dem Stand der Technik den Vorteil, daß keine FFT mit einer vorbestimmten Transformationslänge benötigt wird. Das Verfahren der vorliegenden Erfindung kann somit äußerst sparsam im Hinblick auf die verfügbaren Rechnerressourcen implementiert werden, da das verwendete FIR-Filter einfach zu realisieren ist.

Patentansprüche

1. Verfahren zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal (x(k)), mit folgenden Schritten:
   (a) Segmentieren des zeitdiskreten Audiosignals, um aufeinanderfolgende Segmente gleicher Länge mit ungefilterten zeitdiskreten Audiosignalen ($x_s(T)$, $x_s(T-1)$, $x_s(T-2)$, . . .) zu erzeugen;
   (b) Filtern des zeitdiskreten Audiosignals ($x_s(T)$) in einem aktuellen Segment;
   (c) Vergleichen der Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment mit der Energie ($E_f(T-1)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T-1)$) in einem vorhergehenden Segment; und/oder
   (d) Bestimmen eines aktuellen Verhältnisses zwischen der Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment und der Energie ($E_u(T)$) des ungefilterten zeitdiskreten Audiosignals ($x_s(T)$) in dem aktuellen Segment und Vergleichen des aktuellen Verhältnisses mit einem entsprechenden vorhergehenden Verhältnis; und
   (e) Erfassen eines Anschlags auf der Grundlage des im Schritt (c) und/oder (d) durchgeführten Vergleichs.
2. Verfahren nach Anspruch 1, bei dem der Schritt des Filtern ein Hochpaßfiltern des zeitdiskreten Audiosignals umfaßt.
3. Verfahren nach Anspruch 1 oder 2, bei dem im Schritt (e) ein Anschlag erfaßt wird, wenn der in dem Schritt (c) durchgeführte Vergleich ergibt, daß die Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment größer als die Energie ($E_f(T-1)$) das gefilterten zeitdiskreten Audiosignals ($y_s(T-1)$) in einem vorhergehenden Segment ist.
4. Verfahren nach Anspruch 1 oder 2,
   bei dem in Schritt (c) ferner die Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment mit der Energie ($E_f(T-2)$) eines gefilterten zeitdiskreten Audiosignals ($y_s(T-2)$) in einem zweitletzten vorhergehenden Segment verglichen wird, und
   bei dem im Schritt (e) nur dann ein Anschlag erfaßt wird, wenn die Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment sowohl größer als die Energie ($E_f(T-1)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T-1)$) in dem vorhergehenden Segment als auch größer als die Energie ($E_f(T-2)$) des zeitdiskreten Audiosignals ($y_s(T-2)$) in dem zweitletzten vorhergehenden Segment ist.
5. Verfahren nach Anspruch 1 oder 2,
   bei dem im Schritt (c) ferner die Energie des gefilterten zeitdiskreten Audiosignals im aktuellen Segment mit einem vorbestimmten Filterminimalwert (E_{minF}), der auf der psychoakustischen Ruhehörschwelle basiert, verglichen wird, und
   bei dem im Schritt (c) nur dann ein Anschlag erfaßt wird, wenn die Energie des gefilterten zeitdiskreten Audiosignals in dem aktuellen Segment sowohl größer als die Energie des gefilterten zeitdiskreten Audiosignals in dem vorhergehenden Segment als auch größer als die Energie des gefilterten zeitdiskreten Audiosignals in dem zweitletzten vorhergehenden Segment als auch größer als der vorbestimmte Filter-Minimalwert (E_{minF}) ist.
6. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Energien, die jeweils mit der Energie des gefilterten zeitdiskreten Audiosignals im aktuellen Segment verglichen werden, mit Faktoren (k_1, k_2) gewichtet werden, die größer als eins sind.
7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem im Schritt (e) ein Anschlag erfaßt wird, wenn der in dem Schritt (d) durchgeführte Vergleich ergibt, daß das aktuelle Verhältnis größer als das vorhergehende entsprechende Verhältnis ist.
8. Verfahren nach einem der Ansprüche 1 bis 6,
   bei dem in dem Schritt (e) ferner die Energie ($E_u(T)$) des ungefilterten zeitdiskreten Audiosignals ($x_s(T)$) in dem aktuellen Segment mit einem vorbestimmten Minimalwert (E_{minU}), der auf der psychoakustischen Ruhehörschwelle basiert, verglichen wird, und
   bei dem im Schritt (e) nur dann ein Anschlag erfaßt wird, wenn sowohl das aktuelle Verhältnis größer als das entsprechende vorherige Verhältnis ist, als auch die Energie ($E_u(T)$) des ungefilterten zeitdiskreten Audiosignals ($x_s(T)$) in dem aktuellen Segment größer als der vorbestimmte Minimalwert (E_{minU}) ist.
9. Verfahren nach Anspruch 7 oder 8, bei dem das vorhergehende Verhältnis mit einem vorbestimmten Faktor (k_3), der größer als eins ist, gewichtet wird.

10. Verfahren nach einem der vorhergehenden Ansprüche, bei dem das Hochpaßfiltern mittels eines FIR-Filters durchgeführt wird.

11. Vorrichtung (**12**) zum Erfassen eines Anschlags in einem zeitdiskreten Audiosignal (x(k)) mit folgenden Merkmalen:

(a) einer Einrichtung (**14**) zum Segmentieren des zeitdiskreten Audiosignals (x(k)), um aufeinanderfolgende Segmente mit gleicher Länge mit ungefilterten zeitdiskreten Audiosignalen ($x_s(T)$, $x_s(T-1)$, $x_s(T-2)$, . . .) zu erzeugen;

(b) einem Filter (**16**) zum Filtern des zeitdiskreten Audiosignals ($x_s(T)$) in einem aktuellen Segment;

(c) einer Anstiegserfassungseinrichtung (**20**) zum Vergleichen der Energie ($E_f(T)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment mit der Energie ($E_f(T-1)$) des gefilterten zeitdiskreten Audiosignals ($y_s(T-1)$) in einem vorhergehenden Segment; und/oder

(d) einer Spektralerfassungseinrichtung (**18**) zum Bestimmen eines aktuellen Verhältnisses zwischen der Energie ($E_f(T)$) des ungefilterten zeitdiskreten Audiosignals ($y_s(T)$) in dem aktuellen Segment und der Energie ($E_u(T)$) des gefilterten zeitdiskreten Audiosignals ($x_s(T)$) in dem aktuellen Segment und Vergleichen des aktuellen Verhältnisses mit einem vorhergehenden entsprechenden Verhältnis; und

(e) einer Einrichtung (**22**) zu Erfassen eines Anschlags auf der Grundlage des durch die Anstiegserfassungseinrichtung (**20**) und/oder des durch die Spektralerfassungseinrichtung (**18**) durchgeführten Vergleichs.

12. Vorrichtung (**12**) nach Anspruch 11, bei dem das Filter (**16**) ein Hochpaß-FIR-Filter mit linearem Phasenverhalten ist.

13. Vorrichtung nach Anspruch 11 oder 12, bei dem die Einrichtung (**22**) zum Erfassen eines Anschlags als UND- oder als ODER-Gatter ausgeführt ist, wobei in Eingänge des ODER-Gatters bzw. UND-Gatters Ausgangssignale (kritA, kritB) der Anstiegserfassungseinrichtung (**20**) und der Spektralerfassungseinrichtung (**18**) eingespeist werden.

14. Vorrichtung (**10**) zum Codieren eines zeitdiskreten Audiosignals, mit folgenden Merkmalen:

(a) einer Anschlagserfassungseinrichtung (**12**) zum Erfassen eines Anschlags in dem zeitdiskreten Audiosignal nach einem der Ansprüche 10 bis 12;

(b) einer Einrichtung (**104**) zum Fenstern des zeitdiskreten Audiosignals, um Blöcke von zeitdiskreten Audiosignalen zu erzeugen, die auf die Anschlagserfassungseinrichtung (**12**) anspricht, um ein kurzes Fenster zum Fenstern zu verwenden, wenn die Anschlagserfassungseinrichtung (**12**) einen Anschlag erfaßt;

(c) einer Einrichtung (**104**) zum Zeit/Frequenz-Transformieren der Blöcke des zeitdiskreten Audiosignals, um Blöcke von Spektralkomponenten zu erzeugen; und

(d) einer Einrichtung (**106**) zum Quantisieren und Codieren der Blöcke von Spektralkomponenten.

15. Verfahren zum Codieren eines zeitdiskreten Audiosignals mit folgenden Schritten:

(a) Erfassen eines Anschlags nach einem der Ansprüche 1 bis 9;

(b) Fenstern des zeitdiskreten Audiosignals mit einem kurzen Fenster, wenn ein Anschlag erfaßt wurde, und mit einem langen Fester, wenn kein Anschlag erfaßt wurde, um Blöcke von zeitdiskreten Audiosignalen zu erzeugen;

(c) Transformieren der Blöcke des zeitdiskreten Audiosignals von dem Zeit- in den Frequenzbereich, um Blöcke mit Spektralkomponenten zu erzeugen; und

(d) Quantisieren und Codieren der Blöcke von Spektralkomponenten, um ein codiertes Audiosignal zu erhalten.

Hierzu 2 Seite(n) Zeichnungen

- Leerseite -

Nummer: DE 197 36 669 C1
Int. Cl.[6]: G 10 L 5/00
Veröffentlichungstag: 22. Oktober 1998

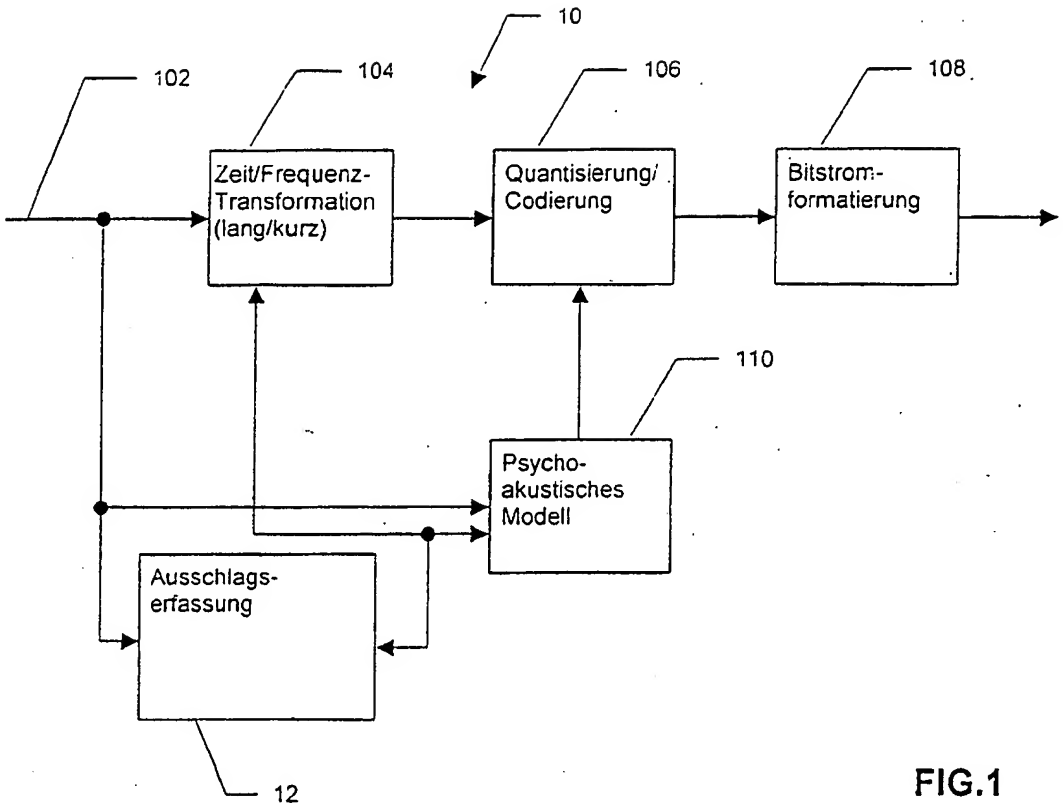

FIG.1

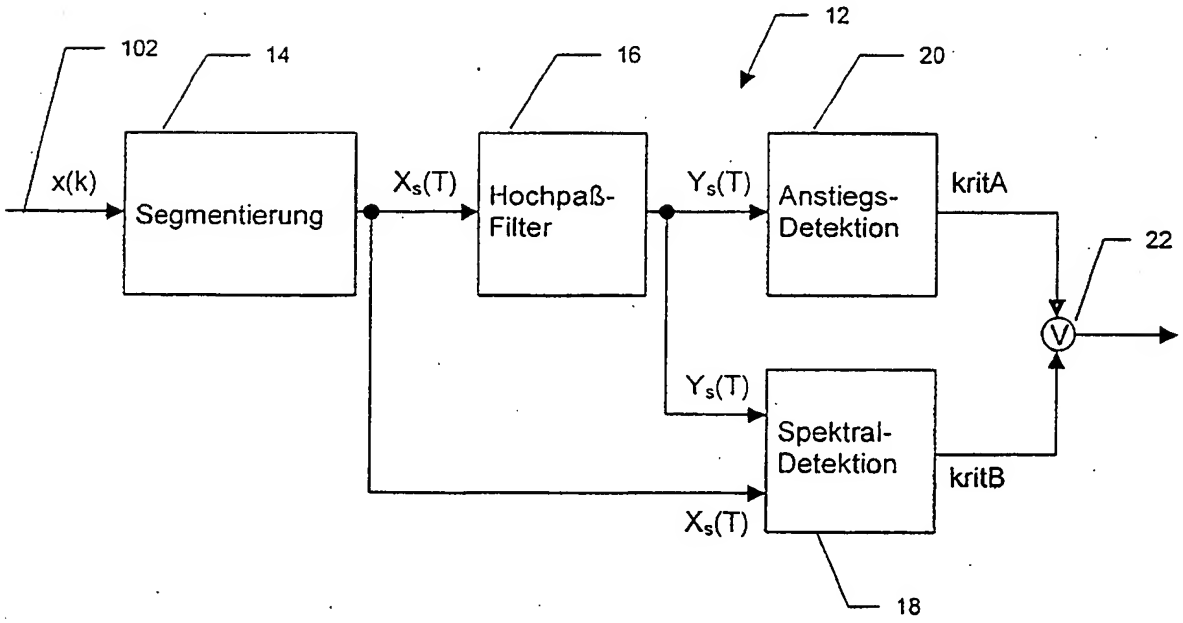

FIG.2

Nummer: DE 197 36 669 C1
Int. Cl.[6]: G 10 L 5/00
Veröffentlichungstag: 22. Oktober 1998

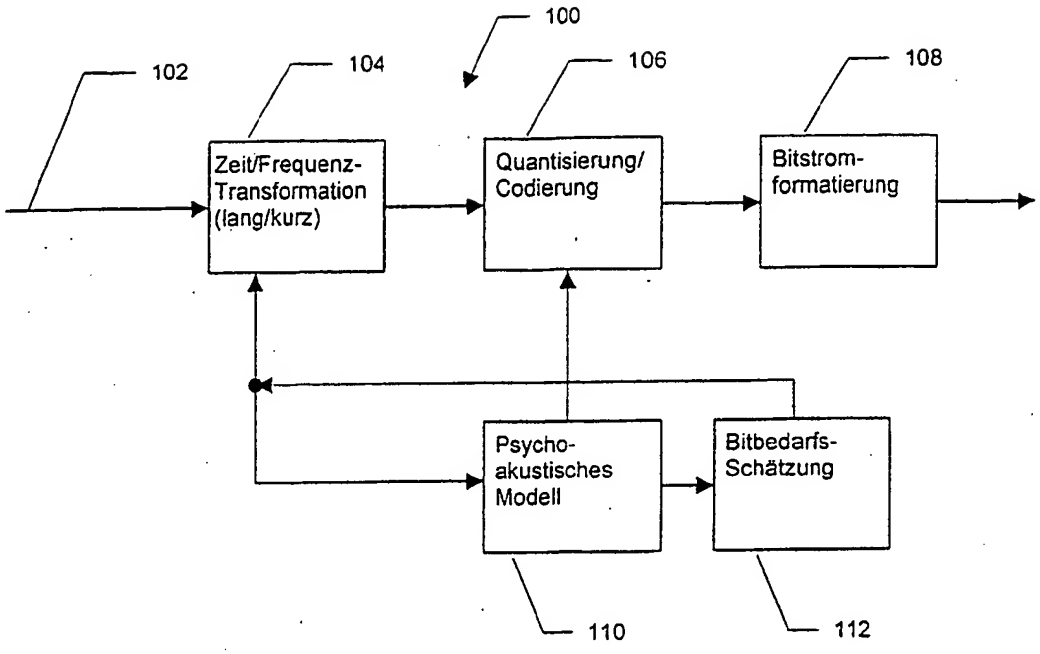

FIG.3 (Stand der Technik)

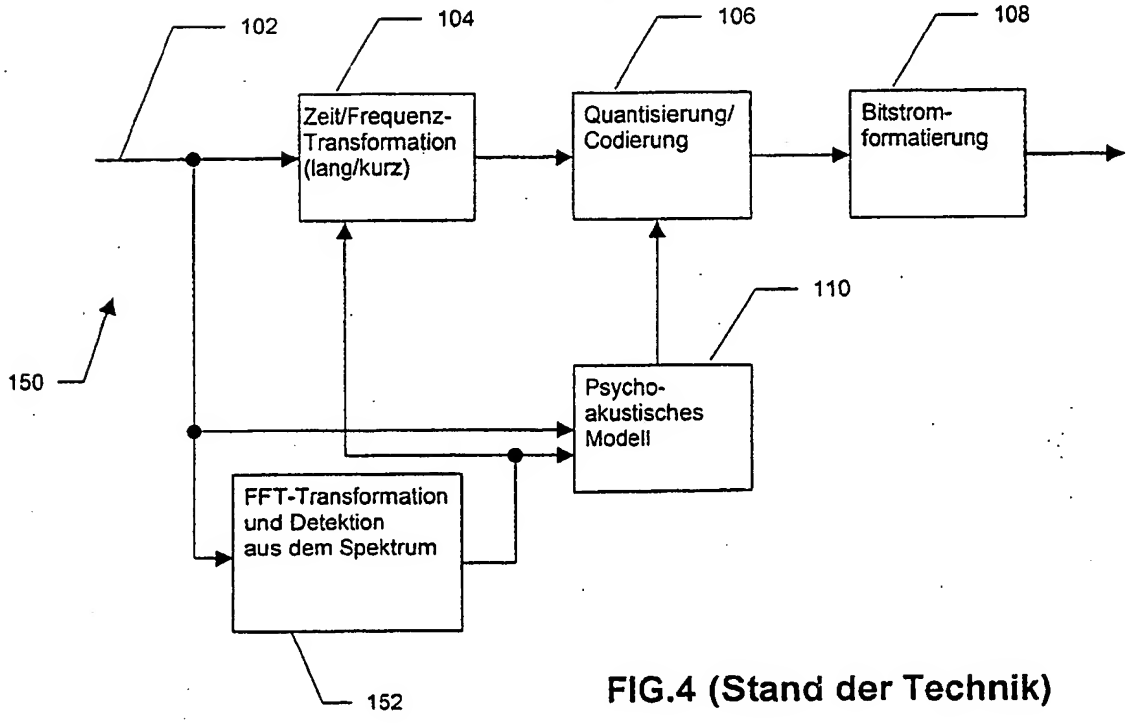

FIG.4 (Stand der Technik)